方差分析作业

2051498储岱泽

研究问题和假设

问题背景

一个在线零售商希望提高员工的工作效率，同时改善他们的工作体验。目前，零售商订单管理中心的员工在工作时没有得到任何形式的娱乐，如背景音乐、电视等。零售商想知道提供一些员工要求的播放背景音乐是否会提高生产力，如果能提高，具体能提高多少。

数据调查

研究人员随机抽取了 150 名员工。这 150 名参与者被随机分为三组，每组50 名参与者：

- **对照组**（control_group）: 不听音乐；
- **治疗组1**（treatment_group_1）:他们听音乐，但不能选择听什么；

- **治疗组2**（treatment_group_2）:他们不仅可以听音乐，还可以自主选择听什么音乐。

在实验结束时，三组的“生产力(productivity)”是根据“每小时处理的平均包裹数量”来衡量。因此，因变量是“生产力”：由一个月实验期间每小时处理的平均包装数量衡量，而解释变量是“分组”：有三个相互独立的组，“no_music”（对照组）、“music_no_choice”（治疗组 1）和“music_choice”（治疗组 2）。

问题假设

研究人员假设，适当的娱乐放松会提高生产力，处于可选择音乐工作状态下的员工（治疗组 2）的生产力水平最高，其次是不可选择音乐工作状态下（治疗组 1），最后是无音乐组（对照组）。

数据分析方法选择

在对这个问题进行分析的过程中，我们选择了**单因素方差分析（One-Way ANOVA）**作为数据分析方法。

选择理由

1. **多个组之间的比较**：我们有三个相互独立的组，即对照组、治疗组A和治疗组B。一元方差分析适用于比较多个组之间的均值差异，并帮助我们判断这些差异是否显著。
2. **检验因素对生产力的影响**：我们想要了解不同处理（无音乐、有音乐但无选择、有音乐且有选择）对员工生产力的影响。一元方差分析可以帮助我们确定因变量（生产力）与解释变量（分组）之间的关系，并检验它们之间是否存在显著差异。
3. **统计显著性检验**：通过一元方差分析，我们可以计算出F统计量和相关的p-value，用来判断不同组之间的生产力差异是否具有统计学上的显著性。

因此，基于以上理由，我们选择了一元方差分析作为数据分析方法，以研究不同组之间的生产力差异并判断其统计学显著性。

基础分析

数据的导入与提取

首先我们导入ANOVA_data_music.csv的数据，并查看数据结构。

```
# 指定数据文件路径
file_path <- "ANOVA_data_music.csv"
# 加载CSV文件
data <- read.csv(file_path)
# 查看数据框的结构
str(data)
# 查看数据的前几行
head(data)
```

```
'data.frame':   150 obs. of  3 variables:
 $ ID          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ condition   : chr  "no_music" "no_music" "no_music" "no_music" ...
 $ productivity: num  188 196 194 190 157 ...
```

其中，ID是指员工的ID，condition标注出了员工在实验中的分组信息，productivity是员工在相应的condition下工作的生产效率。

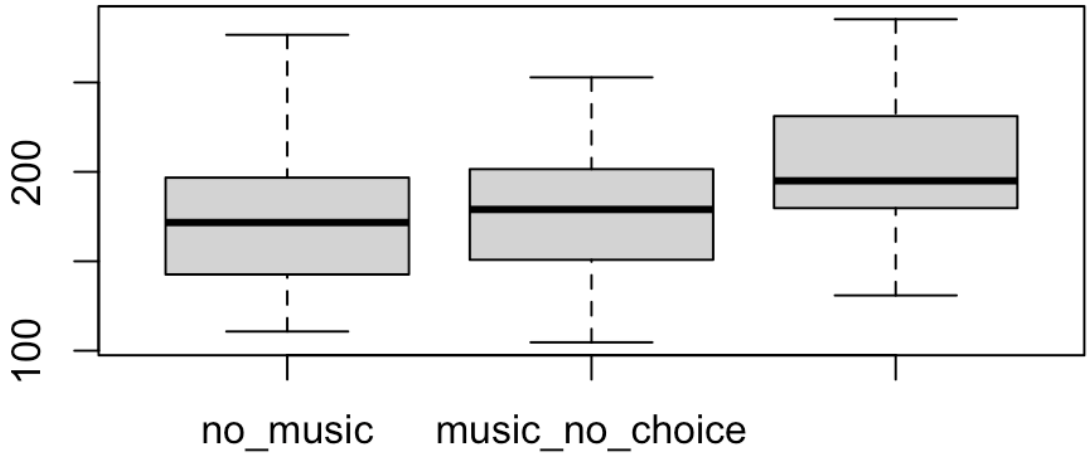
读入数据之后，我们分别提取出control_group，treatment_group_1以treatment_group_2这三个组的数据，并且分别保存。

```
control_group<-subset(data,data$condition=='no_music')
treatment_group_1<-subset(data,data$condition=='music_no_choice')
treatment_group_2<-subset(data,data$condition=='music_choice')
```

数据分布分析

我们通过对三组数据分别绘制箱线图来研究三组数据的分布情况：

```
# 将数据整合成一个列表
data_list <- list(control_group$productivity,
                  treatment_group_1$productivity,
                  treatment_group_2$productivity)
# 画箱线图
boxplot(data_list, names=c('no_music', 'music_no_choice', 'music_choice'))
```



同时输出最小值、下四分位数、中位数、上四分位数、最大值。

```
# 使用stats输出最小值、下四分位数、中位数、上四分位数、最大值以及离群点
result <- boxplot(data_list, plot=FALSE)
result$stats  # 输出最小值、下四分位数、中位数、上四分位数、最大值
result$out    # 输出离群点
```

```
         [,1]     [,2]     [,3]
[1,] 110.7401 104.6698 130.9050
[2,] 142.6391 150.8186 179.7325
[3,] 171.6782 178.9807 194.9928
[4,] 196.7702 201.5226 231.1774
[5,] 276.6144 252.8472 285.3483
numeric(0)
```

所以我们可以得到：

	no_music	music_no_choice	music_choice
最小值	110.7401	104.6698	130.9050
下四分位数	142.6391	150.8186	179.7325
中位数	171.6782	178.9807	194.9928

	no_music	music_no_choice	music_choice
上四分位数	196.7702	201.5226	231.1774
最大值	276.6144	252.8472	285.3483

结合图表我们可以初步分析得出，就中位数而言，no_music的组工作效率最低，music_no_choice的组工作效率略高，music_choice的组工作效率显然比前两组高。

同时，虽然就下四分位数和上四分位数以及中位数而言，music_no_music的组都有更高的效率，但是no_music组的最小值和最大值都更高，因此不能显著区分这两组的工作效率，差不多，不过music_choice的组的工作效率显著比前两组高。

主要分析

方差分析

对数据的分布和一些基础数值有了了解之后，接下来我们对数据进行单因素方差分析。

```
# 进行方差分析
anova_result <- aov(productivity ~ condition, data = data)
summary(anova_result)
```

输出结果：

```
             Df Sum Sq Mean Sq F value   Pr(>F)
condition     2  24734   12367   9.291 0.000159 ***
Residuals   147 195661    1331
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

根据方差分析的结果，我们可以看到**自变量“condition”对因变量的影响是显著的**。具体来说，F值为9.291，对应的p值为0.000159，远小于一般的显著性水平（比如0.05）。这表示我们可以接受原假设，即是否能够自主选择听音乐确实对员工的工作效率有显著影响。因此，我们可以得出结论，是否能够自主选择听音乐确实对员工的工作效率有显著影响。

两两之间比较：no_music与music_no_choice

```
# 执行独立样本t检验
result <- t.test(control_group$productivity, treatment_group_1$productivity)
# 输出结果
print(result)
```

```
        Welch Two Sample t-test

data:  control_group$productivity and treatment_group_1$productivity
t = -0.36286, df = 94.389, p-value = 0.7175
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.12278  11.83114
sample estimates:
mean of x mean of y
 174.4983  177.1442
```

- t值为-0.36286，表示两组之间的**工作效率差异相对较小**。
- p值为0.7175，大于通常的显著性水平（例如0.05），这意味着在统计学上没有足够的证据拒绝原假设。原假设可以理解为**两组之间的工作效率没有显著差异**。
- 样本估计值显示控制组的平均工作效率为174.4983，处理组1的平均工作效率为177.1442。

总结来说，根据这次t检验的结果，在我们可以看见不听音乐工作和听不能自己选择的音乐工作没有显著差别。

两两之间比较：no_music与music_choice

```
# 执行独立样本t检验
result <- t.test(control_group$productivity, treatment_group_2$productivity)
# 输出结果
print(result)
```

```
        Welch Two Sample t-test

data:  control_group$productivity and treatment_group_2$productivity
t = -3.7225, df = 97.263, p-value = 0.0003305
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -43.64312 -13.28968
sample estimates:
mean of x mean of y
 174.4983  202.9647
```

- t值为-3.7225，表示"no_music"组和"music_choice"组的工作效率之间存在**显著差异。**
- p值为0.0003305，远小于通常的显著性水平（例如0.05），这意味着在统计学上有足够的证据拒绝原假设。原假设可以理解为"no_music"组和"music_choice"组的工作效率没有显著差异。
- 样本估计值显示"no_music"组的平均工作效率为174.4983，"music_choice"组的平均工作效率为202.9647。

总结来说，根据这次t检验的结果，在统计学上可以得出结论，即**"no_music"组和"music_choice"组的工作效率存在显著差异。**

两两之间比较：music_no_choice与music_choice

```
# 执行独立样本t检验
result <- t.test(treatment_group_1$productivity, treatment_group_2$productivity)
# 输出结果
print(result)
```

```
        Welch Two Sample t-test

data:  treatment_group_1$productivity and
treatment_group_2$productivity
t = -3.7239, df = 96.819, p-value = 0.0003297
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -39.58260 -12.05856
sample estimates:
mean of x mean of y
 177.1442  202.9647
```

- t值为 -3.7239，表明在 **“music_no_choice” 和 “music_choice” 的工作效率之间存在显著差异。**
- p值为 0.0003297，远小于通常的显著性水平（例如0.05），这意味着在统计学上有充分的证据拒绝原假设。原假设可以理解为 treatment_group_1 和 treatment_group_2 的工作效率没有显著差异。
- 样本估计值显示 treatment_group_1 的平均工作效率为 177.1442，treatment_group_2 的平均工作效率为 202.9647。

综合来看，根据这次 t 检验的结果，在统计学上可以得出结论，即 **treatment_group_1 和 treatment_group_2 的工作效率存在显著差异。**

分析结果解读

因此，结合方差分析的结果，以及箱线图的情况来看，三组之间的生产力存在统计学上的显著差异。

- “no_music”和”music_no_choice”的组生产力没有明显区别。
- “no_music”和”music_choice”的组生产力有明显区别。
- “music_no_choice”和”music_choice”的组生产力有明显区别。

同时，”music_choice”的组具有最高的生产力水平。所以在员工工作的时候播放他们自己选择的音乐可以提升工作效率。